Direct estimation of the minimum RSS value for training Bayesian Knowledge Tracing parameters

Francesc Martori
ASISTEMBE
IQS Universitat Ramon Llull
Barcelona, Spain
francesc.martori@iqs.edu

Jordi Cuadros
ASISTEMBE
IQS Universitat Ramon Llull
Barcelona Spain
jordi.cuadros@iqs.edu

Lucinio González-Sabaté
ASISTEMBE
IQS Universitat Ramon Llull
Barcelona, Spain
lucinio.gonzalez@iqs.edu

ABSTRACT
Student modeling can help guide the behavior of a cognitive tutor system and provide insight to researchers on understanding how students learn. In this context, Bayesian Knowledge Tracing (BKT) is one of the most popular knowledge inference models due to its predictive accuracy, interpretability and ability to infer student knowledge. However, the most popular methods for training the parameters of BKT have some problems, such as identifiability, local minima, degenerate parameters and computational cost during fitting. In this paper we address some of the issues of one of these training models, BKT Brute Force. Instead of finding the parameter values that provide the lowest Residual Sum of Squares (RSS), we estimate this minimum RSS value from some a priori known values of the skill. From there we perform some preliminary analysis to improve our knowledge of the relationship between the RSS, from BKT-BF, and the four BKT parameters.

Keywords: Bayesian Knowledge Tracing · BKT Brute Force · RSS modeling

1. INTRODUCTION

1.1 Bayesian Knowledge Tracing
Bayesian Knowledge Tracing (BKT) [1] is a student model used to infer a student's knowledge given their history of responses to problems, which it can use to predict future performance. Using students' responses to questions, which are tagged with the skills that the instructor wants the students to learn, the model tells the probability a student has mastered a skill.

BKT is a two state Hidden Markov Model, these states being the one in which the student knows a given skill, and the one where the student does not. The "knowledge" state is absorbent, implying that the student will not forget the skill once it is learned. To calculate the probability that a student knows the skill given their performance history, BKT needs to know four probabilities:

L_0, the probability a student knows the skill before attempting the first problem,

T, the probability a student, who does not currently know the skill, will know it after the next practice opportunity, that is the transition probability at each practice opportunity,

G, the probability a student will answer a question correctly despite not knowing the skill,

S, the probability a student will answer a question incorrectly despite knowing the skill.

According to this model, knowledge affects performance (mediated by the guess and slip rates), and knowledge at one time step affects knowledge at the next time step: if a student is in the unknown state at time t, then the probability they will be in the "knowledge" state at time t+1 is P(T). Usually, a separate BKT model is fit for each skill and only the first attempt at each question is taken for each student.

1.2 Bayesian Knowledge Tracing – Brute Force
Bayesian Knowledge Tracing – Brute Force [2] (BKT-BF) is an algorithm to estimate the values for the BKT parameters. It is a simple brute force algorithm, where a grid of possible values is set so that for each combination of parameters, a RSS value is obtained. At the end, the combination of values resulting in the lowest Residual Sum of Squares (RSS) value for a skill is the one that will be used in BKT.
In BKT-BF, the RSS is calculated as follows:

$$RSS = \sum_{i}^{students} \sum_{t=1}^{dim} \left(O_{i,t} - C_{i,t}\right)^2 \quad \text{eq. 1}$$

Where:

$O_{i,t}$ is $\{0,1\}$ depending on the student's answer to a given question,

students is the number of different students who faced any question of a given skill,

dim is the number of different questions that are tagged with a given skill

$C_{i,j}$ is the likelihood to produce a correct answer to a question. This calculation is derived from the BKT formulas, and it is done, for the student i, as follows:

$$C_{i,t} = L_{t-1} * (1 - S) + (1 - L_{t-1}) * G \quad \text{eq. 2}$$

BKT-BF is, however, is very expensive in computational cost, as all brute force algorithms are, and does not help the identifiability [3] problem from BKT; identifiability results in different combinations of parameter values, some of which make no theoretical sense, giving similar RSS values. The other most usual algorithm is EM [4], which is not as computationally demanding but suffers from local minima issues. There are efforts to develop methods [5], [6], [7] and [8] that use different techniques to tackle the issues we mentioned, however, in this paper we will focus our work on BKT-BF.

Given that BKT-BF is an algorithm that gives good practical results, but it is so computationally expensive, the objective of this paper is to make accurate estimates of the minimum RSS value for any skill. At the same time, this might provide a better understanding of the BKT model.

2. DATA AND METHODS

The data used belongs to the 'Psychology MOOC GT - Spring 2013' dataset, accessed via DataShop (pslcdatashop.org) [9]. This course was designed by the Open Learning Initiative (OLI), who are known for their data driven design [10], [11], this fact and their long experience in course design ensure that skills have been properly tagged. The course was taken by 5615 students that issued around 2 million first attempt answers. There were 226 different skills identified in the course. The skills map used can be also found in [9]

In order to obtain the RSS values, we have used the BKT-BF algorithm. Specifically, we have used values from 0.05 to 0.95, with a 0.15 step, for L_0 and T; for G and S, the bounded approach has been taken in order to avoid model degeneracy [5], so we have used values from 0.05 to 0.30, with a 0.05 step. Given all this, 1764 different RSS values were obtained per skill.

To identify each skill, we have defined three variables:

- *dim*: number of different questions that are tagged with a given skill
- *n*: total number of responses on questions tagged with a given skill. It's the product of *students* and *dim* from eq.1
- *percent_correct (pc)*: Percentage of correct answers to questions tagged with a given skill

These variables have been chosen as they are pieces of information that one may have easy access to before computing BKT-BF.

In order to achieve the aforementioned objective, we will train a linear model using the three variables we defined for each skill. This model will allow us to make predictions of which will be any skill's minimum RSS, if we were to train it using BKT-BF. To train the model, we have extracted the minimum RSS value, resulting from the BKT-BF calculations, for each one of the skills, and used it as the RSS value for that skill. An example of the data we have worked with is shown in table 1.

It has to be noted, that skills that were tagged in less than 4 different questions (*dim*<4) have been discarded. That results in a sample of 103 different skills for training and evaluating the model.

Table 1. Data structure for skills

dim	n	pc	Skill	Grid BKT-BF	RSS min
8	4923	79,8%	1	1764 data	795.6
16	11062	89,5%	2	1764 data	1024.5
…	…	…	…	…	…

The resulting distribution of RSS values is far from being normal, as it could be expected. However, if instead of using the RSS value, we compute the Root Mean Squared Error (RMSE) for each skill, by taking the square root of the RSS divided by n, the resulting distribution is acceptably normal as we can see in the histogram shown in Figure 1 and in the Q-Q plot in Figure 2. This latter plot assesses normality by displaying the normal theoretical quantiles (x axis) and the normal data quantiles (y axis). If the distribution is perfectly normal, data would perfectly fit the dotted line.

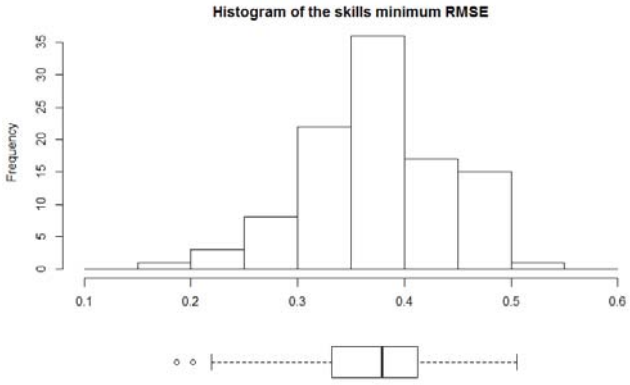

Figure 1. Histogram and boxplot of the RMSE distribution

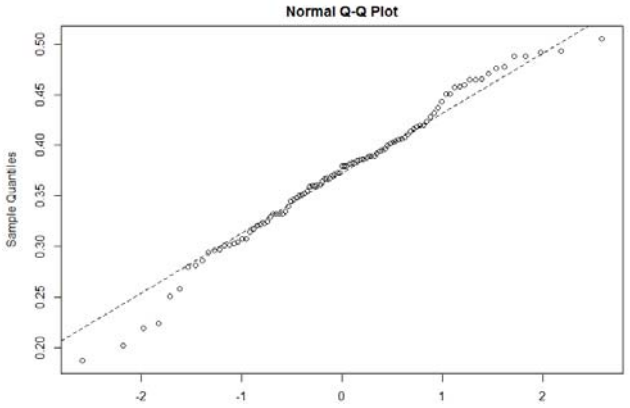

Figure 2. Q-Q plot of the RMSE

3. RESULTS

Firstly, a brief summary for the data we have worked with is shown in the table 2.

Table 2. Summary of the data for training the model.

	n	dim	pc	RMSE	RSS
min	1738	4.00	0.458	0.187	152.0
Median	5899	8.00	0.821	0.379	811.8
Mean	8556	9.65	0.807	0.373	1235.1
Max	47215	23.00	0.964	0.505	8483.4

A linear regression has been performed on the RMSE, using *n*, *dim*, *pc*, some usual transformations, such as using the logarithms and the squares of the variables, and the variable interactions as predictors. A best subset selection (using the leaps package in R, [12]) approach has been taken, resulting the best model the one using a second degree polynomial with *pc*. The results for the linear regression estimates are shown in the table 3.

Table 3. Linear regression results and error metrics

Variable	Estimate	Std. Error	t value	P(>|t|)
Intercept	0.3725	0.0009	415.9	<2e-16
pc	-0.6096	0.0091	-67.1	<2e-16
pc^2	-0.2545	0.0091	-28.0	<2e-16
Adjusted R^2		**Residual standard Error**		
0.981		0.009089		

Finally, using a random validation set (75 skills to train the model and 28 to test it), we have obtained an adjusted R^2 of 0.978, that shows a very good predictive ability for the adjusted model.

In an attempt to have a better knowledge on the relationship between the *RMSE* values and the BKT parameters, we have run a preliminary Principal Components Analysis (PCA). The resulting biplot of the PCA is shown in figure 3. For the sake of a proper understanding of the relationship between the different variables, we have eliminated the data labels from the chart. The variance explained by the first two Components of PCA is 71.4%.

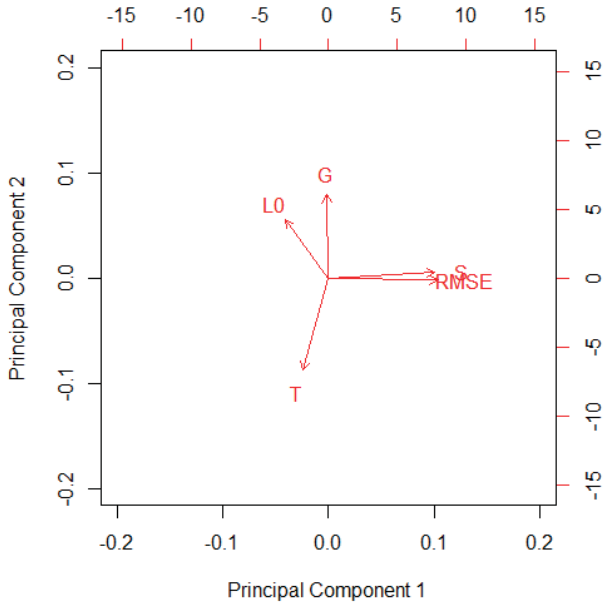

Figure 3. Plot of the PCA loadings of RMSE, L_0, T, G and S

In the chart, we can see how the RMSE is highly correlated with the slip parameter. At the same time, the parameters *G* and *T* seem to be highly inversely correlated, which is something that one can expect as the more likely it is to learn a skill, the less likely it is that you might be guessing the outcome. However, the most noticeable aspect is the orthogonality between *T*, *G* and *RMSE*. In the PCA context, orthogonality is related to poorly correlated variables. If that was to be true, it could imply that T and G have little or no effect in terms of RMSE variation. We have also calculated and drawn the biplots for each skills' RMSEs, using all BKT-BF data points, not just the minimums, and their results lead us to similar conclusions than the ones obtained from figure 3.

4. DISCUSSION AND CONCLUSIONS

We have been able to find a linear model that allows us to estimate the minimum RSS value for the training of the BKT parameters. Using this, we might be able to find a quicker convergence using a modified version of BKT-BF, so that the computational cost will be reduced. Even though that the model has been developed using the RMSE instead of the RSS, the model will also be useful for predicting the latter as the only difference is a transformation involving *dim* and *n*.

We are aware that, in the BKT-BF calculations, we are using a step much larger than the one recommended by the algorithm. This shouldn't be a problem with the conclusions we reached because we are not using BKT-BF for estimating the BKT parameters, but to generate data with which we train a model for estimating the minimum RSS for any skill.

The very high performance of the model, in terms of adjusted R^2, may be indicating that BKT works better when the percentage of correct answers is very high, as the RSS decreases. This has some implications in the BKT model because if the percentage of correct answers is very high, there might not be much room for T and G in the model. We would only be trying to adjust the probability of already knowing the skills before doing the course and the probability of slipping.

To be more certain about the conclusions stated here, the following steps have to include using, at least, a different dataset to shed some light around the suspicions that arise on the influence of T and G in the BKT model. A deeper analysis beyond an exploratory PCA is also required.

5. AKNOWLEDGEMENTS

We acknowledge the help received from colleagues at OLI at Stanford University.

6. REFERENCES

[1] Corbett, A. T. and Anderson, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278.

[2] Baker, Corbett, Gowda, Wagner, MacLaren, Kauffman, Mitchell, & Giguere, 2010. Bayesian Knowledge Tracing Brute Force model fitting code. http://users.wpi.edu/~rsbaker/edmtools.html

[3] Beck, J. E., Chang, K. M. 2007 Identifiability: A fundamental problem of student modeling. In: Conati, C., McCoy, K., Paliouras, G. (Eds.) UM 2007. LNCS, vol. 4511/2007, pp. 137- 146.

[4] Moon, T. K. 1996. The expectation–maximization algorithm. IEEE Signal Process. Mag., 13, 47–60

[5] Baker, R.S.J.d., Corbett, A. T., Aleven, V.: More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In: Woolf, B., Aimeur, E., Nkambou, R., Lajoie, S. (Eds.) ITS 2008. LNCS, vol. 5091/2008, pp. 406-415. Springer, Berlin Heidelberg (2008)

[6] Hawkins, W., Heffernan, N.T., Baker, R.S.J.d. (2014) Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. Lecture Notes in Computer Science Volume 8474, 2014, pp 150-155.

[7] Pardos, Z. A., Heffernan, N. T.: Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In: Baker, R.S.J.d., Merceron, A., Pavlik, P.I. (Eds.) Proceedings of the 3rd International Conference on Educational Data Mining, pp. 161-170 (2010)

[8] Rai, D., Gong, Y., Beck, J.: Using Dirichlet priors to improve model parameter plausibility. In: Barnes, T., Desmarais, M., Romero, C., Ventura, S. (Eds.) Proceedings of the 2nd International Conference on Educational Data Mining, pp. 141-150 (2009)

[9] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. 2010. A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

[10] Strader, R., Thille, C.: The Open Learning Initiative: Enacting Instruction Online. In Oblinger D. G. (Ed.), Game Changers. Education and Information Technologies, pp. 201-213 (2012)

[11] Thille, C. (2012). Changing the Production Function in Higher Education. Making Productivity Real. American Council on Education 2012.

[12] Thomas Lumley using Fortran code by Alan Miller (2009). leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/package=leaps